

特開平5-233873

(43)公開日 平成5年(1993)9月10日

(51)Int.Cl.⁵

G 0 6 K 9/20

// G 0 6 F 15/70

識別記号

3 4 0 L

3 3 0 Q 9071-5L

庁内整理番号

F I

技術表示箇所

審査請求 未請求 請求項の数12(全 32 頁)

(21)出願番号 特願平4-267313

(22)出願日 平成4年(1992)10月6日

(31)優先権主張番号 特願平3-341889

(32)優先日 平3(1991)11月29日

(33)優先権主張国 日本(JP)

(71)出願人 000006747

株式会社リコー

東京都大田区中馬込1丁目3番6号

(72)発明者 立川 道義

東京都大田区中馬込1丁目3番6号 株式
会社リコー内

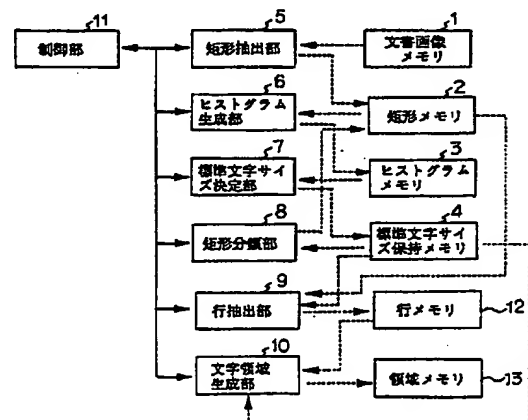
(74)代理人 弁理士 鈴木 誠 (外1名)

(54)【発明の名称】 領域分割方法

(57)【要約】

【目的】 文字サイズの違う様々な文書に対して的確な領域分割を行なう。

【構成】 矩形抽出部5により黒連結成分の外接矩形を抽出し、ヒストグラム生成部6で矩形の高さのヒストグラムを生成し、これにもとづき標準文字サイズ決定部7が標準文字サイズを決定する。矩形分類部8は、標準文字サイズと矩形の大きさを比較し、文字矩形と図表矩形进行分类する。行抽出部9は文字矩形を統合した行を抽出し、文字領域生成部10は行を統合した文字領域を抽出する。



【特許請求の範囲】

【請求項1】 文書画像より黒連結成分に外接した矩形を抽出し、抽出した矩形の高さのヒストグラムより標準文字サイズを決定し、抽出した矩形を、その大きさと標準文字サイズとの大小関係に基づいて、文字の矩形とそれ以外の矩形とに分類することを特徴とする領域分割方法。

【請求項2】 文書画像より黒連結成分に外接した矩形を抽出し、抽出した矩形の高さのヒストグラムより標準文字サイズを決定し、抽出した矩形を、その大きさと標準文字サイズとの大小関係に基づいて、文字の矩形とそれ以外の矩形とに分類し、文字の矩形を統合し文字列の行を抽出することを特徴とする領域分割方法。

【請求項3】 文書画像より黒連結成分に外接した矩形を抽出し、抽出した矩形の高さのヒストグラムより標準文字サイズを決定し、抽出した矩形を、その大きさと標準文字サイズとの大小関係に基づいて、文字の矩形とそれ以外の矩形とに分類し、文字の矩形を統合し文字列の行を抽出し、抽出した行を統合して行の集合である文字領域を抽出することを特徴とする領域分割方法。

【請求項4】 請求項1, 2または3記載の領域分割方法において、文書画像の処理対象領域の境界からの距離がある閾値以下の文字以外の矩形をノイズとして除去することを特徴とする領域分割方法。

【請求項5】 請求項4記載の領域分割方法において、閾値を標準文字サイズに応じて変化させることを特徴とする領域分割方法。

【請求項6】 文書画像より、文字に対応した黒連結成分に外接した矩形を抽出し、抽出した文字の矩形の統合により行を生成する領域分割方法において、文書画像より垂直罫線を抽出し、ある注目した矩形より矩形の統合を行なう際に、当該注目矩形との水平距離が水平方向の統合条件である矩形間距離の閾値より小さく、かつ、当該注目矩形と垂直方向の重なりを持つ垂直罫線と遭遇した場合、当該注目領域に関し、水平方向の統合条件である矩形間距離の閾値を、当該注目矩形と当該垂直罫線との水平距離に対応した値に変更することを特徴とする領域分割方法。

【請求項7】 文書画像より、文字に対応した黒連結成分に外接した矩形を抽出し、抽出した文字の矩形の統合により行を生成し、生成した行の統合により文字領域を生成する領域分割方法において、文書画像より垂直罫線を抽出し、ある注目した矩形より矩形の統合を行なう際に、当該注目矩形との水平距離が水平方向の統合条件である矩形間距離の閾値より小さく、かつ、当該注目矩形と垂直方向の重なりを持つ垂直罫線と遭遇した場合、当該注目領域に関し、水平方向の統合条件である矩形間距離の閾値を、当該注目矩形と当該垂直罫線との水平距離に対応した値に変更し、

ある注目した行より行の統合を行なう際に、統合しようとする行が、その生成時に垂直罫線との遭遇により水平統合条件たる矩形間距離の閾値を変更したものである場合、当該統合しようとする行の水平方向の終点座標より当該注目行の水平方向の終点座標が大きいたときには、当該統合しようとする行の統合を禁止することを特徴とする領域分割方法。

【請求項8】 請求項7記載の領域分割方法において、行統合により生成された文字領域について注目した文字領域より文字領域の再統合を行ない、その際に、統合しようとする文字領域内の行の中に、その生成時に垂直罫線との遭遇により水平方向の統合条件たる矩形間距離の閾値を変更した行が含まれている場合、当該統合しようとする文字領域の水平方向の終点座標より当該注目文字領域の水平方向の終点座標が大きいたときには、当該統合しようとする文字領域の統合を禁止することを特徴とする領域分割方法。

【請求項9】 文書画像より、文字に対応した黒連結成分に外接した矩形を抽出し、抽出した文字の矩形の統合により行を生成し、生成した行の統合により文字領域を生成する領域分割方法において、文書画像より水平罫線を抽出し、ある注目した行より行の統合を行なう際に、当該注目行との垂直距離が垂直方向の統合条件である行間距離の閾値より小さく、かつ、当該注目行と水平方向の重なりを持つ水平罫線と遭遇した場合、当該注目行に関し、垂直方向の統合条件である行間距離の閾値を、当該注目行と当該水平罫線との垂直距離に対応した値に変更することを特徴とする領域分割方法。

【請求項10】 請求項9記載の領域分割方法において、行統合により生成された文字領域について注目した文字領域より文字領域の再統合を行ない、その際に、統合しようとする文字領域内の行の中に、その生成時に水平罫線との遭遇により垂直方向の統合条件たる行間距離の閾値を変更した行が含まれている場合、当該統合しようとする文字領域の垂直方向の終点座標より当該注目文字領域の垂直方向の終点座標が大きいたときには、当該統合しようとする文字領域の統合を禁止することを特徴とする領域分割方法。

【請求項11】 文書画像より、文字に対応した黒連結成分に外接した矩形を抽出し、抽出した文字の矩形の統合により文字領域を生成する領域分割方法において、文書画像より文字以外の大きな黒連結成分に外接した矩形を抽出し、抽出した文字以外の矩形を水平方向にスキャンして閾値以上の長い黒ランのみからなる黒連結成分に外接する矩形Hを抽出し、

それぞれの文字以外の矩形より抽出された矩形Hの中で、当該文字以外の矩形の上辺または下辺から所定距離範囲内で最も上または最も下に位置し、かつ水平罫線と

しての形状条件を満足したものを架空水平セパレータとして抽出し、

文字の矩形の統合による文字領域の生成の際に、架空水平セパレータを文字領域を上下に区切る水平セパレータとして扱うことを特徴とする領域分割方法。

【請求項12】 文書画像より、文字に対応した黒連結成分に外接した矩形を抽出し、抽出した文字の矩形の統合により文字領域を生成する領域分割方法において、文書画像より文字以外の大きな黒連結成分に外接した矩形を抽出し、

抽出した文字以外の矩形を垂直方向にスキャンして閾値以上の長い黒ランのみからなる黒連結成分に外接する矩形Vを抽出し、

それぞれの文字以外の矩形より抽出された矩形Vの中で、当該文字以外の矩形の左辺または右辺から所定距離範囲内で最も左または最も右に位置し、かつ垂直罫線としての形状条件を満足したものを架空垂直セパレータとして抽出し、

文字の矩形の統合による文字領域の生成の際に、架空垂直セパレータを文字領域を左右に区切る垂直セパレータとして扱うことを特徴とする領域分割方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、図表と文字などが混在した一般文書の画像に対し、文字の領域と、図領域や表領域等の文字以外の領域（図・表領域）とを識別分類し、必要な領域分割を行なうための領域分割方法に関する。

【0002】

【従来の技術】 文書画像を光ディスク等の記憶メディアにファイリングする場合、ファクシミリで通信する場合、プリンタに出力する場合等に、文書画像を文字領域と図・表領域に切り分け、文字領域には文字領域用処理を、図・表領域には図の処理や表の処理を施したいことがある。光学的文字認識装置を用いて文書上の文字をコード化する場合にも、同様の文字領域と図・表領域の切り分けが必要になる。

【0003】 従来、このような領域抽出に関する技術として、原稿画像から黒連結成分の外接矩形を抽出し、抽出した矩形の大きさを予め定めた閾値と比較することによって、文字の矩形と線図形の矩形を判定する方法が知られている（特開昭55-162177号）。

【0004】

【発明が解決しようとする課題】 しかし、黒連結成分の外接矩形の種類判定に予め定められた閾値を用いるので、文字サイズが異なる様々な文書の画像が入力された場合に柔軟に対応できない。

【0005】 領域分割の対象となる原稿が書籍や雑誌のページである場合、1ページずつ切り取ってスキャナにセットすることは稀で、書籍や雑誌の目的ページを開

き、見開き原稿としてスキャナにセットし読み取らせることが多い。このような見開き原稿を読み取った画像の場合、原稿の中折れ部つまりページ間境界部に黒ずんだ、あるいは黒い領域が生じやすい。また、スキャナの押さえ板（スキャナの光源の光を反射させ画像の地肌を白にするためのもの）を閉じることができず、あるいは、これが浮き上がり、その結果、原稿の周辺部に黒領域が生じやすい。このような見開き原稿画像の領域分割を行なう場合、中折れ部や周辺部に観測される黒領域をノイズ領域として適切に処理すべきであるが、従来は、このノイズ処理を考慮していなかった。

【0006】 文書画像においては、罫線を間に文字領域（コラム）が接近していることがあるが、従来は、このような文字領域のセパレータとしての罫線に対する考慮が払われていないため、罫線により区切られた隣接文字領域が一つの文字領域に統合されてしまったり、表領域の文字領域やグラフ領域内の文字領域が、それに隣接した外部の文字領域と統合されてしまうことがあった。

【0007】 本発明は、より高精度に文字領域の分離を行なうための領域分割方法を提供しようとするものである。より具体的に言えば、本発明の一つの目的は、文字サイズ等が様々な文書に対して、文字の領域とそれ以外の領域とを的確に分離し、必要な領域分割を行なう領域分割方法を提供することである。もう一つの目的は、書籍のような見開き原稿や厚手の原稿より読み取った文書画像に対して、中折れ部や周辺部に生じやすいノイズ領域の影響を排除し、的確な領域分割が可能な領域分割方法を提供することにある。本発明の他の目的は、罫線により区切られた文字領域や、表領域あるいはグラフ領域の内部の文字領域を、的確に分離することができる領域分割方法を提供することにある。

【0008】

【課題を解決するための手段】 請求項1乃至5の発明によれば、文書画像より抽出した黒連結成分に外接した矩形の高さのヒストグラムより標準文字サイズを決定し、抽出した矩形を、標準文字サイズとの大小関係に基づいて、文字の矩形とそれ以外の矩形とに分類し、そのうえで文字の矩形のみを統合することによって文字領域を生成する。

【0009】 請求項4または5の発明によれば、文書画像の処理対象領域の境界からの距離が、予め決められた閾値または標準文字サイズに応じて決められた閾値以下の文字以外の矩形をノイズとして除去する。

【0010】 請求項6乃至8の発明によれば、文書画像より文字の矩形を抽出するほかに垂直罫線を抽出し、ある注目した矩形より矩形の統合を行なう際に、当該注目矩形との水平距離が水平方向の統合条件である矩形間距離の閾値より小さく、かつ、当該注目矩形と垂直方向の重なりを持つ垂直罫線と遭遇した場合、当該注目領域に関し、水平方向の統合条件である矩形間距離の閾値を、

10

20

30

40

50

当該注目矩形と当該垂直罫線との水平距離に対応した値に変更する。

【0011】請求項7の発明によれば、文字の矩形の統合により生成された行の統合を行なうが、注目した行より行を統合する際に、統合しようとする行が、その生成時に垂直罫線との遭遇により水平統合条件たる矩形間距離の閾値を変更したものである場合、その行の水平方向の終点座標より当該注目行の水平方向の終点座標が大きいときには統合を禁止する。

【0012】請求項8の発明によれば、行の統合により生成された文字領域について再統合を行なうが、ある注目した文字領域より文字領域の統合を行なう時に、統合しようとする文字領域内の行の中に、その生成時に垂直罫線との遭遇により水平方向の統合条件たる矩形間距離の閾値を変更した行が含まれている場合、当該統合しようとする文字領域の水平方向の終点座標より当該注目文字領域の水平方向の終点座標が大きいときには統合を禁止する。

【0013】請求項9または10の発明によれば、文書画像より文字の矩形を抽出するほかに垂直罫線を抽出し、文字の矩形を統合した行について、ある注目した行より行統合を行なう際に、当該注目行との垂直距離が垂直方向の統合条件である行間距離の閾値より小さく、かつ、当該注目行と水平方向の重なりを持つ水平罫線と遭遇した場合、当該注目行に関し、垂直方向の統合条件である行間距離の閾値を、当該注目行と当該水平罫線との垂直距離に対応した値に変更する。

【0014】請求項10の発明によれば、行統合により生成された文字領域について注目した文字領域より再統合を行なうが、その際に、統合しようとする文字領域内の行の中に、その生成時に水平罫線との遭遇により垂直方向の統合条件たる行間距離の閾値を変更した行が含まれている場合、当該統合しようとする文字領域の垂直方向の終点座標より当該注目文字領域の垂直方向の終点座標が大きいときには統合を禁止する。

【0015】請求項11または12の発明によれば、文書画像より、文字の矩形を抽出するほかに、文字以外の大きな図・表矩形を抽出する。請求項11の発明によれば、図・表矩形を水平方向にスキャンして閾値以上の長い黒ランのみからなる黒連結成分に外接する矩形Hを抽出し、それぞれの文字以外の矩形より抽出された矩形Hの中で、当該文字以外の矩形の上辺または下辺から所定距離範囲内で最も上または最も下に位置し、かつ水平罫線としての形状条件を満足したものを架空水平セパレータとして抽出し、この架空水平セパレータを文字領域を上下に区切る水平セパレータとして扱って、文字の矩形を統合する。請求項12の発明によれば、図・表矩形を垂直方向にスキャンして閾値以上の長い黒ランのみからなる黒連結成分に外接する矩形Vを抽出し、それぞれの文字以外の矩形より抽出された矩形Vの中で、当該

文字以外の矩形の左辺または右辺から所定距離範囲内で最も左または最も右に位置し、かつ垂直罫線としての形状条件を満足したものを架空垂直セパレータとして抽出し、この架空垂直セパレータを文字領域を左右に区切る垂直セパレータとして扱って、文字の矩形の統合を行なう。

【0016】

【作用】請求項1乃至5の発明によれば、様々な文字サイズの文書の画像に対して、文字の矩形と、それ以外の矩形（図表矩形）とを正確に分類できるようになるため、文字矩形の集合たる文字領域の抽出精度が向上する。

【0017】請求項4または5の発明によれば、書籍や厚手の原稿から読み取られた文書画像に対しても、中折れ部や周辺部に生じる黒領域の影響を排除し、文字矩形、行または文字領域を正確に抽出することができる。特に請求項5の発明によれば、処理しようとする文書の文字サイズの違いに柔軟に対応したノイズ矩形除去を行なうことができる。

【0018】請求項6乃至8の発明によれば、垂直セパレータで区切られた文字領域（コラム）の誤統合を防止し、それぞれの文字領域を正しく分離できる。

【0019】請求項9または10の発明によれば、水平セパレータで区切られた文字領域の誤統合を防止し、それぞれの文字領域を正しく分離できる。

【0020】請求項11または12の発明によれば、表やグラフの領域と、それに接近した文字領域との誤統合を防止し、文字領域を正しく抽出できる。

【0021】以下、本発明の特徴と利点をより明確にするため、図面を用い様々な実施例について説明する。

【0022】

【実施例】図1は、後記実施例1乃至実施例6を説明するためのブロック図である。

【0023】1はスキャナによって読み取られた文書画像を格納するための文書画像メモリである。この文書画像の処理の対象となる領域は、文書画像全体としてもよいし、あるいは、文書画像をディスプレイ画面に表示しマウス等を用いて指定した領域としてもよい。

【0024】5は入力された文書画像の黒連結成分に外接した矩形を抽出する矩形抽出部である。ただし、矩形抽出部5の前段に画像圧縮手段を設け、入力文書画像を圧縮した画像（例えば8×8画素を1画素に圧縮した画像）に対し、矩形抽出を行なってもよい。2は抽出された矩形の情報を記憶するための矩形メモリである。

【0025】6は矩形の高さのヒストグラムを生成するヒストグラム生成部、3はこのヒストグラムの情報を記憶するためのヒストグラムメモリ、7は矩形の高さのヒストグラムより標準文字サイズを決定する標準文字サイズ決定部、4は決定した標準文字サイズの情報を記憶するための標準文字サイズ保持メモリである。

【0026】8は抽出された矩形を、文字の矩形とそれ以外の矩形(図表の矩形)に分類する矩形分類部である。9は文字矩形を統合して文字列の行を抽出する行抽出部、12は抽出された行の情報を記憶するための行メモリである。10は行を統合して行の集合である文字領域を生成する文字領域生成部、13は生成された文字領域の情報を記憶するための領域メモリである。11は前記各部の動作を制御する制御部である。

【0027】なお、各処理部及び制御部11は、ハードウェアで実現しても、コンピュータシステム上でソフトウェアにより実現しても、あるいはハードウェアとソフトウェアの組み合わせにより実現してもよい。いずれの実現形態をとるかは、必要とされる処理速度等を考慮して選択すればよい。

【0028】図2は処理概要の説明図である。(a)に示す原稿画像が入力した場合に、矩形抽出処理によって(b)に示す矩形データが得られる。(d)は矩形データを拡大して示す。この矩形データに対し、矩形統合と行統合を行なうことによって(c)に示す行/領域データが得られる。以下、このような処理の内容について実施例毎に説明する。

【0029】実施例1

図2は矩形抽出から標準文字サイズ決定までの処理フローを示す。ステップ100は矩形抽出部5の処理で、入力された文書画像上の黒連結を抽出し、それに外接する矩形を抽出し、例えば矩形の左上角と右下角の座標と矩形番号などを矩形情報として矩形メモリ2に格納する。

【0030】ステップ105はヒストグラム生成部6の処理で、矩形メモリ2内の矩形情報をもとに矩形の高さ(行に対して垂直方向の矩形サイズである。縦書き文書の場合、矩形の横方向のサイズである)のヒストグラムを生成する。このヒストグラムの例を図4に示す。

【0031】ステップ110~115は、標準文字サイズ決定部7の処理である。本実施例の場合、矩形の高さのヒストグラムの総度数(矩形数)を T とし、 $t = T/16$ を算出する(ステップ110)。次に、矩形の高さのヒストグラムをもとに、度数が t を超える矩形高さの中で最大の矩形高さを当該文書の標準文字サイズ $size$ とする(ステップ111~115)。したがって、図4の(a)に示すヒストグラムが得られた場合には $size = 4$ に決定され、(b)に示すヒストグラムが得られた場合には $size = 6$ に決定される。

【0032】次に、矩形分類部8によって、標準文字サイズ $size$ を基準に用いて矩形分類を行なう。例えば、高さ及び幅(行方向のサイズ)のいずれもが標準文字サイズ $size$ の5倍を超えない矩形を文字の矩形として分類し、それ以上に大きな矩形を文字以外の矩形(図・表矩形)とする。

【0033】この分類結果が本実施例の出力である。すなわち、黒連結成分の矩形を単位として文書画像を文字

領域と図・表領域とに分割する。図2で言えば、同図(b)の矩形データが本実施例の処理結果である。

【0034】実施例2

矩形分類部8による矩形分類までの処理は前記実施例1と同一であるが、次に行抽出部9によって次のような抽出を行なう。

【0035】二つの文字矩形の距離(行方向)を求め、この距離が標準文字サイズ $size$ 以下であれば両文字矩形を統合する。横書き文書の場合、例えば図5に示す文字矩形151、152の水平距離 $sp1$ を求め、 $sp1 \leq size$ ならばそれらを統合する。縦書き文書ならば、縦方向の距離により同様の統合を行なう。このような統合処理を全部の文字矩形に対して行ない、統合された文字矩形群の外接矩形を文字列の行として切り出す。図6において、153が切り出された行であり、その内部の矩形が文字矩形である。

【0036】このように本実施例は、文字矩形を統合し、文書画像の文字領域を行を単位として抽出するわけである。図2で言えば、(b)に示した矩形データに対し文字矩形の統合により行を抽出した段階のデータが、処理結果である。

【0037】実施例3

矩形分類の後に文字矩形を統合して行を抽出することは前記実施例2と同様であるが、文字矩形の統合方法が異なる。

【0038】すなわち、横書き文書の場合、例えば図6の文字矩形161、162の重なり量 $sp2$ を測定し、矩形高さ $h1$ 、 $h2$ の小さいほうの値を h として、 $sp2 \geq h/2$ ならば文字矩形161、162を統合する。このようにして統合した文字矩形群の外接矩形を行として抽出する。

【0039】縦書き文書の場合、縦方向についての文字矩形間の重なり量と、文字矩形の高さ(横方向の大きさ)を用いて同様の判定を行ない統合して、行を抽出する。

【0040】実施例4

行抽出部9で文字矩形の統合による行抽出を行なうが、この際に、前記実施例2における文字矩形の統合条件と、前記実施例3における文字矩形の統合条件の両方を満足した場合に、文字矩形の統合を行なう。

【0041】実施例5

行抽出部9により行が抽出されるまでの処理は前記実施例2、3または4と同様である。

【0042】次に文字領域生成部10において、次に述べるような行の統合を行なって文字領域(コラム)を生成する。まず、 p に初期値として $size$ を設定し、統合しようとする行との距離 $p2$ が p 以下のときは、その行の統合を行なうが、 $p2$ が p を越えるときは統合を行なわない。

【0043】行を統合した場合には、統合した行の間の

距離 p_1 (行に対し垂直方向の距離) の1.5倍の値 p を求める (p_1 が一定値以下、例えば1以下の場合、 p を固定値、例えば3に設定する)。そして、次に統合しようとする行との距離 p_2 が p 以下のときは、その行を統合し、 p を超えるときは統合しない。

【0044】ただし、 $p_2 \leq p$ であっても、統合した最後の行の高さ h_1 と次に統合しようとする行の高さ h_2 の比 h_2/h_1 が、ある定数 (例えば4/5) 以上の場合は統合しない。

【0045】図8は、1番目の統合条件の説明図である。ただし横書き文書の場合である。行171, 172を統合し、次に行173について $p_2 \leq p$ であるため、これを統合する。これによって、文字領域174が得られる。

【0046】図9は、2番目の統合条件の説明図である。行181, 182が統合され、次に行183を統合する際に、 $h_2/h_1 \geq 4/5$ であるため統合しない。その結果、行181, 182の文字領域186が得られる。この例は、文字領域186のほかに、行183のみの文字領域187と、行184, 185を統合した文字領域188が生成された場合である。

【0047】このように本実施例によれば、文書画像より、行がまとまった文字領域を抽出することができる。図2で言えば、(c)の行/領域データが得られる。

【0048】実施例6

文字領域生成部10は、前記実施例5と同様の行統合による文字領域生成を行なうが、その統合の完璧を期するため、さらに次のような文字領域に対する統合処理を行なう。

【0049】統合しようとする二つの文字領域の幅が、両方とも標準文字サイズ $size$ の5倍以上の場合と、少なくとも一方の幅が標準文字サイズ $size$ の5倍より小さい場合とで、統合条件が異なる。

【0050】図10は、前者の場合、つまり大きい領域同士の統合の場合の説明図である。191, 192は実施例5と同様の統合処理によって生成された文字領域である。文字領域191, 192の行方向の重なり量を sp_1 、文字領域191の幅を w 、文字領域191, 192の距離を sp_2 とし、 $sp_1 \geq 4/5w$ であって、 $sp_2 < size$ の場合、文字領域191, 192を一つ

の文字領域193に統合する。

【0051】図11は、後者の小さい領域の統合の場合の説明図である。195, 196は統合しようとする文字領域である。文字領域195, 196の行方向の距離を sp_1 、小さいほうの文字領域196の高さを h とし、文字領域195, 196の行に対し垂直方向の重なり量が $h/2$ 以上で、かつ $sp_1 \leq 2 \times size$ のとき、一つの文字領域197に統合する。

【0052】実施例7

図12は本実施例及び後記実施例8, 9を説明するため

のブロック図である。ノイズ矩形除去部201が追加されていることが、図1と異なる。

【0053】文書画像メモリ1に入力された文書画像の原稿が、書籍や厚手の原稿であった場合に、原稿の中折れ部や周辺部に黒領域もしくは黒ずんだ領域が生じることが前述した。本実施例は、このような黒領域に対応した図・表矩形 (文字矩形以外の矩形) をノイズとして除去するもので、そのために設けられた手段がノイズ矩形除去部201である。

【0054】図13は、ノイズ矩形除去部201の処理フロー図である。矩形メモリ2を参照し、矩形分類部8によって分類後の矩形の情報を読み出し、図・表矩形 (文字矩形以外の矩形) であるか調べ (ステップ210)、図・表矩形であれば、処理対象領域のいずれかの境界辺に接しているか否かを調べる (ステップ215)。ここで、処理対象領域は、前述のように、文書原稿の全体または指定された領域である。境界辺に接触している場合、その図・表矩形をノイズ矩形として除去する。除去された文字矩形は、次の行抽出の対象矩形から除外される。

【0055】実施例8

本実施例においてもノイズ矩形除去を行なうことは前記実施例7と同様であるが、ノイズ矩形の判定方法が多少異なる。すなわち、図14に示すように、処理対象領域230の境界の各辺と、図・表矩形との水平または垂直距離が、予め定められた閾値 $NOISM$ 以下であった場合、その矩形をノイズ矩形として除去する。

【0056】図14の例では、図・表矩形231, 232は除去される。ただし、文字矩形233は除去しない。なお、 $NOISMG=0$ に設定した場合、本実施例と前記実施例7のノイズ矩形の判定除去は実質的に同一となる。

【0057】実施例9

本実施例は前記実施例8と同様のノイズ矩形除去を行なうが、閾値 $NOISMG$ として、標準文字サイズ $size$ の整数倍の値を用いることが違う。このようにすることで、ノイズ領域の幅を入力文書の文字サイズに応じ適応的に変化させることができる。

【0058】図15は、後記実施例10乃至14を説明するためのブロック図である。301は入力された文書画像データを記憶するための文書画像メモリ、306は入力文書画像の黒連結成分の外接矩形を抽出する矩形抽出部、302は抽出された矩形の情報を記憶する矩形メモリである。

【0059】矩形抽出の処理は、前記実施例1~9と同様でよく、例えば文書画像をスキャンしながら接続した黒欄の外接矩形を抽出し、これを一定距離内にあるものについて統合することによって行なう。なお、矩形抽出部306の前に入力文書画像の例えば 8×8 画素を1画素に圧縮 (縮小) する手段を設け、圧縮された画像に対

10

20

30

40

50

して矩形抽出を行なうようにしてもよい。

【0060】307は矩形を文字矩形と、それ以外の図・表矩形に分類する矩形分類部である。この矩形分類の処理は、前記実施例1乃至9と同様に、矩形の幅と高さを閾値と比較することによって行なうが、その際の閾値は前記実施例1乃至9と違って予め与えられる。ただし、前記実施例1乃至9と同様に、例えば矩形の高さのヒストグラムから標準文字サイズを求め、これに適当な係数を掛けることによって適応的に閾値を算出してもよい。

【0061】308は、垂直罫線（垂直セパレータ）を識別する垂直罫線識別部である。この垂直罫線識別は、例えば次のような処理によって行なわれる。矩形分類部307の処理によって分類された図・表矩形（文字矩形以外の大きな矩形）の中で、その幅（横方向の大きさ）が所定の閾値より小さく、かつ、高さ（縦方向の大きさ）が所定の閾値より大きい矩形を垂直罫線の候補として選ぶ。次に、垂直罫線候補矩形の範囲の文書画像を垂直方向にスキャンすることにより、所定の閾値より長い黒ランのみを抽出し、この長い黒ランのみの連結成分の外接矩形を抽出する。そして、この長い黒ラン連結成分の外接矩形の長さ $H1$ 、幅 $W1$ と、もとの矩形の高さ H 、幅 W との間で、 $H1/H > \text{閾値}$ （例えば0.8）かつ $W1/W > \text{閾値}$ （例えば0.8）であれば、長い黒ラン連結成分の外接矩形を垂直罫線であると判定する。

【0062】なお、垂直罫線矩形の判定のための閾値、及び黒ランの長さの閾値は、矩形分類の閾値と同様に適応的に決定してもよい。特願平4-160866号の明細書及び図面に、このような罫線抽出の方法がより詳細に示されている。

【0063】309は文字矩形を統合して行を生成する行抽出部であり、303は行生成の際の文字矩形の統合判定のための閾値を記憶する統合閾値メモリである。304は生成された行の情報を記憶する行メモリである。310は生成された行を統合し文字領域を生成する文字領域生成部である。305は生成された文字領域の情報を記憶するための領域メモリである。行統合の処理は、前記実施例1乃至9と同様に、統合しようとする行の間の距離と統合閾値とを比較することによって統合可否を判定し、可能な行間を統合するという方法で行なわれるが、この統合判定の際に、垂直罫線識別部308により抽出された垂直罫線の有無と相対的位置関係が参照される。311は文字領域生成部310の行統合によって生成された文字領域に対して、統合し切れなかった領域の再統合を行ない文字領域を修正する文字領域修正部である。312は前記各部を制御する制御部である。

【0064】以下、実施例10乃至14について個別に説明する。ただし、文書上の行は横方向であると仮定する。

【0065】実施例10

矩形抽出部306での矩形抽出処理、矩形分類部307による矩形分類処理、垂直罫線識別部308による垂直罫線抽出に続いて、行抽出部309により行生成が行なわれる。

【0066】図16は行生成処理の概略フローを示す。行生成の基本的な処理は、文字矩形と判定された矩形に対し、ある注目した矩形より、他の矩形との間の水平距離（ $sp1$ ）及び垂直距離（ $sp2$ ）と、予め標準文字サイズ等から定められて統合閾値メモリ303に格納されている水平統合閾値（ $Th1$ ）及び垂直統合閾値（ $Th2$ ）とを比較することで、近接した矩形を統合するというものであるが、図16に見られるように、垂直罫線との関係に応じて水平統合閾値を変更する。

【0067】まず、新しい文字矩形Aを選択し（ステップ360）、水平統合閾値を $Th1$ とする（ステップ362）。注目した文字矩形Aより右方向へ、水平距離が $Th1$ 以内かつ垂直距離が垂直統合閾値 $Th2$ 以内の範囲で矩形（文字矩形または垂直罫線矩形）をサーチし、見つかった矩形が垂直罫線矩形であるか判定する（ステップ364）。その矩形が文字矩形であるならば統合し（ステップ366）、次の矩形のサーチを行なう。

【0068】文字矩形の統合を目的としているのであるから、サーチされた矩形が垂直罫線矩形である場合には、当然にその垂直罫線矩形との統合を行なわない。そして、注目している矩形Aと垂直罫線矩形とが垂直方向で重なりがあるか調べ（ステップ368）、重なりがある場合には、矩形Aと垂直罫線矩形の中心線までの距離 $Th1'$ を水平統合閾値に設定し直し（ステップ370）、次の矩形のサーチを行なう。

【0069】矩形Aに関して、水平統合閾値（ $Th1$ または $Th1'$ ）以下かつ垂直統合閾値 $Th2$ 以下の範囲内に次の矩形が見つからなくなるまで処理を繰り返すと（ステップ372）、ステップ360より別の注目矩形に関して処理を開始する。

【0070】文字矩形の統合中に垂直罫線と遭遇する場合の具体例を、図17により説明する。図17(a)において、矩形Aに関して、まず矩形Bが統合される。次に矩形Aから $Th1$ 以下の水平距離内に垂直罫線矩形380が見つかる。この垂直罫線矩形380は矩形Aと重なりがあるので、水平統合閾値が $Th1$ から $Th1'$ へ変更され、その結果、水平統合閾値が $Th1$ であれば次にサーチされ統合されるはずであった矩形Cは矩形Aとの統合対象から除外される。よって、矩形A、Bからなる行381と、垂直罫線（垂直セパレータ）により区切られた矩形Cとが正しく分離される。

【0071】図17(b)において、矩形Aに対して矩形Bが統合され、次に垂直罫線矩形382がサーチされる。しかし、この垂直罫線矩形382は矩形Aとの重なりがないので、水平統合閾値は $Th1$ のままである。したがって、次の矩形Cがサーチされて矩形Aと統合さ

10

20

30

40

50

れ、矩形A、B、Cからなる行383が生成される。すなわち、この垂直罫線矩形382は、少なくとも矩形Bと矩形Cとの間の垂直セパレータではないので、その存在を無視して文字矩形の統合を行なう。

【0072】実施例11

前記実施例10と同様に、行生成処理において、注目する文字矩形と重なりのある垂直罫線矩形が存在した場合に水平統合閾値を変更するが、その変更する値の決定方法が異なる。これ以外は前記実施例10と同様である。

【0073】図18は、水平統合閾値の変更値の決定方法の説明図である。矩形Aの垂直方向の範囲(Ya1からYa2)と、垂直罫線矩形384の水平方向の範囲(Xr1からXr2)で決まるスキャン範囲385に対して、画像を水平方向にスキャンし、あるスキャンライン上の垂直罫線の存在位置(Xr3)を求め、距離(Xr3-Xa2)を変更後の水平統合閾値とする。あるいは、全スキャンラインについて求めた垂直罫線の位置の平均値をXr3として、(Xr3-Xra)を変更後の水平統合閾値としてもよい。

【0074】実施例12

前記実施例10と同様に、行生成処理において、注目する文字矩形と重なりのある垂直罫線矩形が存在した場合に水平統合閾値を変更するが、その変更する値の決定方法が異なる。これ以外は前記実施例10と同様である。

【0075】図19は、水平統合閾値の変更値の決定方法の説明図である。図19(a)及び(b)において、 θ は予め測定した文書画像のスキュー角度である。このスキュー角度の測定は、よく知られたハフ変換等の方法によって求めてもよいし、あるいは同一行に属する文字矩形の最大高さとの高さの差より算出する方法などによってもよい。

【0076】(a)は $\theta \geq 0$ の場合であり、矩形Aの対角頂点座標と垂直罫線矩形386の対角頂点座標より、次式

$$Th1' = (Yr2 - (Ya1 + Ya2)/2) \tan \theta + Xr1 - Xa2$$

によって水平統合閾値の変更値Th1'を計算する。

【0077】(b)は $\theta < 0$ の場合であり、矩形Aの対角頂点座標と垂直罫線矩形387の対角頂点座標より、次式

$$Th1' = ((Ya1 + Ya2)/2 - Yr1) \tan \theta + Xr1 - Xa2$$

によって水平統合閾値の変更値Th1'を計算する。

【0078】実施例13

文字矩形の統合による行生成の処理までは前記実施例10乃至12と同様であるが、次に文字領域生成部310により行を統合して文字領域を生成する。この行統合による文字領域生成処理のフローを図20に示す。

【0079】図20の最初のステップ400は、行生成の段階での処理である。このステップでは、注目矩形Aに対して一定距離範囲内で右側に垂直罫線矩形が存在する場合(つまり、前記実施例10乃至12で行生成の際

に水平統合閾値を変更する原因となった垂直罫線が存在する場合)に、生成した行に「右側に垂直罫線が存在する」ということを示すマークRを付加する(ステップ400)。例えば図17(a)の行381にマークRが付けられる。

【0080】ステップ402以降が、行統合による文字領域生成の処理である。まず、新しい行Aを選択する(ステップ402)。この注目した行Aとの水平距離及び垂直距離が一定値以下の行をサーチし、見つかったならば、その行がマークR付きの行であるか調べる(ステップ404)。マークR付き行でなければ、行の統合を行ない(ステップ406)、次の行のサーチを行なうことになる。

【0081】サーチした行がマークR付き行である場合、この行の水平方向の終点座標(Xr2)と行Aの水平方向の終点座標(Xa2)とが、 $Xa2 > Xr2$ の関係を満たすか判定する(ステップ412)。この関係を満たさないときは行統合を行なう(ステップ406)が、関係を満たすときは行統合を行なわず次の行のサーチに進む。

【0082】ステップ412の条件が満たされるために行統合が行なわれない例を図21に示す。図21の(a)において、行413はマークR付きであるので、注目行A414とは統合されない。このような行413、414は、垂直罫線416によって区切られた別々のコラムに属する行と看做することができる。

【0083】図21の(b)において、マークR付き行417と注目行A418とは統合されない。このような行417、418は、別コラムに属するものでない可能性もあるが、垂直罫線419との位置関係から判断すると、統合しないのが自然である。

【0084】実施例14

前記実施例13と同様に、行統合による文字領域生成まで行なうが、統合された行にマークRが付加されている文字領域にもマークRを付加する。そして、行統合により生成された文字領域に対して、文字領域修正部311で行統合と同様な処理により統合を行なうことによって、行統合では統合しきれなかった領域の再統合を行なう。この再統合の際、統合しようとする文字領域がマークR付きであるか調べ、マークR付きの場合は前記実施例13の行統合の際と同様に、マークR付き領域の水平方向の終点座標Xr2と、注目している文字領域の水平方向の終点座標Xa2とが、 $Xa2 > Xr2$ の関係にあるときには統合しない(図21参照)。

【0085】図22は、後記実施例15乃至18を説明するためのブロック図である。501は入力された文書画像データを記憶するための文書画像メモリ、506は入力文書画像の黒連結成分の外接矩形を抽出する矩形抽出部、502は抽出された矩形の情報を記憶する矩形メモリである。矩形抽出の処理は、入力文書画像を圧縮し

た画像に対して行なうようにしてもよい。507は矩形を文字矩形と、それ以外の図・表矩形に分類する矩形分類部である。この矩形分類の処理は、前記実施例10乃至14または前記実施例1乃至9と同様でよい。

【0086】808は、水平罫線（水平セパレータ）を識別する水平罫線識別部である。この水平罫線識別は、例えば次のような処理によって行なわれる。矩形分類部507の処理によって図・表矩形（文字矩形以外の矩形）と分類された矩形の中で、その高さ（縦方向の大きさ）が所定の閾値より小さく、かつ、幅（縦方向の大きさ）が所定の閾値より大きい矩形を水平罫線の候補として選ぶ。次に、水平罫線候補矩形の範囲の文書画像を水平方向にスキャンすることにより、所定の閾値より長い黒ランのみを抽出し、この長い黒ランのみの連結成分の外接矩形を抽出する。そして、この長い黒ラン連結成分の外接矩形の高さ $H1$ 、幅 $W1$ と、もとの矩形の高さ H 、幅 W との間で、 $H1/H > \text{閾値}$ （例えば0.8）かつ $W1/W > \text{閾値}$ （例えば0.8）であれば、長い黒ラン連結成分の外接矩形を水平罫線であると判定する。なお、水平罫線矩形の判定のための閾値、及び黒ランの長さの閾値は、矩形分類の閾値と同様に適応的に決定してもよい。

【0087】509は文字矩形の水平及び垂直方向の距離に関し統合閾値によって統合判定を行ない、統合条件を満たす文字矩形を統合して行を生成する行抽出部である。ただし、水平罫線識別部508で抽出された水平罫線の矩形も、架空の行として抽出する。統合閾値は予め与えられるか、あるいは矩形の高さのヒストグラム等に基づいて適応的に決定される。503は抽出された行の情報を記憶する行メモリである。

【0088】510は生成された行を統合し文字領域を生成する文字領域生成部である。505は生成された文字領域の情報を記憶するための領域メモリである。行統合の処理は、統合しようとする行間の距離と統合閾値とを比較することによって統合可否を判定し、可能な行間を統合するという方法で行なわれるが、この統合判定の際に水平罫線識別部508により抽出された水平罫線の存在に応じて統合閾値が変更される。504は行統合のための統合閾値が格納される統合閾値メモリである。511は文字領域生成部510の行統合によって生成された文字領域に対して、統合し切れなかった領域の再統合を行ない文字領域を修正する文字領域修正部である。512は前記各部を制御する制御部である。

【0089】以下、実施例15乃至18について個別に説明する。ただし、文書上の行は横方向であると仮定する。

【0090】実施例15

矩形抽出部506での矩形抽出、矩形分類部507による矩形分類、水平罫線識別部508による垂直罫線抽出、行抽出部509による行抽出に続いて、文字領域生

成部510による文字領域生成が行なわれる。

【0091】図23は文字領域生成処理の概略フローを示す。文字領域生成の基本的な処理は、注目した行と他の行間の水平距離（ $p1$ ）及び垂直距離（ $p2$ ）と、予め標準文字サイズ等から定められて統合閾値メモリ303に格納されている水平統合閾値（ $Th1$ ）及び垂直統合閾値（ $Th2$ ）とを比較し、統合閾値より近接した矩形を統合するというものであるが、図23に示すように、水平罫線たる架空行に関連して垂直統合閾値の変更操作が行なわれる。

【0092】すなわち、新しい行Aを選択し（ステップ520）、垂直統合閾値を $Th2$ とする（ステップ522）。注目した行Aより下へ向かって、垂直距離が垂直統合閾値 $Th2$ 以内かつ水平距離が水平統合閾値 $Th1$ 以内の範囲で、行（水平罫線矩形も架空行として含める）をサーチし、見つかった行が水平罫線であるか判定する（ステップ524）。その行が本来の行である場合には、注目行Aと統合し（ステップ526）、次の行のサーチを行なう。

【0093】サーチされた行が水平罫線（架空行）である場合には、当然にその水平罫線との統合を行なわない。そして、注目している行Aと水平罫線とが水平方向に重なりを持っているか調べ（ステップ530）、重なりがある場合には、注目している行Aから水平罫線の中心線までの距離 $Th2'$ を、垂直統合閾値に設定し直し（ステップ532）、次の矩形のサーチを行なう。

【0094】行Aに関して、垂直統合閾値（ $Th1$ または $Th1'$ ）以下の垂直距離範囲内に次の行が見つからなくなるまで処理を繰り返すと（ステップ526）、ステップ520より別の注目行に関して処理を開始する。

【0095】水平罫線が存在する場合の具体例を図24により説明する。図24の（a）において、行Aに関する行統合中に水平罫線542と遭遇した場合、この水平罫線542は行Aと重なりがあるので、垂直統合閾値は $Th2$ から $Th2'$ へ変更される結果、垂直統合閾値が $Th2$ であれば次にサーチされ統合されるはずであった行B543は、行Aと統合されなくなる。よって、行Aを含む文字領域544と、水平罫線（水平セパレータ）542により区切られた行B543とが正しく分離される。

【0096】図24の（b）において、行Aに関して水平罫線546がサーチされる。しかし、この水平罫線546は、水平方向について注目行A540と重なりを持たないため、垂直統合閾値は $Th2$ のままである。したがって、次の行B543がサーチされて行A540と統合され、行A540及び行B543を含む文字領域548が生成される。すなわち、この水平罫線546は、少なくとも行A540と行B543との間の水平セパレータではないので、その存在を無視して行統合が行なわれる。

【0097】実施例16

前記実施例15と同様に、文字領域生成（行統合）の処理において、注目する行と重なりのある水平罫線が存在した場合に垂直統合閾値を変更するが、その変更する値の決定方法が異なる。これ以外は前記実施例15と同様である。

【0098】図25は、垂直統合閾値の変更値の決定方法の説明図である。注目している行A550の水平方向の範囲（Xa1からXa2）と、水平罫線の矩形551の垂直方向の範囲（Yr1からYr2）で決まるスキャン範囲552に対して、画像を垂直方向にスキャンし、あるスキャンライン上の水平罫線の存在位置（Yr3）を求め、距離（Yr3-Ya2）を変更後の垂直統合閾値とする。あるいは、全スキャンラインについて求めた水平罫線の位置の平均値をYr3として、（Yr3-Yra）を変更後の垂直統合閾値としてもよい。

【0099】実施例17

前記実施例15と同様に、文字領域成処理において、注目する行Aと重なりのある水平罫線が存在した場合に垂直統合閾値を変更するが、その変更する値の決定方法が異なる。これ以外は前記実施例15と同様である。

【0100】図26は、垂直統合閾値の変更値の決定方法の説明図である。図26（a）及び（b）において、 θ は予め測定した文書画像のスキュー角度である。このスキュー角度の測定は、前記実施例12において述べたような方法で行なえばよい。

（a）は $\theta \geq 0$ の場合であり、注目矩形A553の対角頂点座標と水平罫線の矩形554の対角頂点座標より、次式

$$Th2' = \{ (Xa1 + Xa2) / 2 - Xr1 \} \tan \theta + Yr1 - Ya2$$
によって垂直統合閾値の変更値 $Th2'$ を計算する。

【0101】（b）は $\theta < 0$ の場合であり、矩形A553の対角頂点座標と水平罫線矩形555の対角頂点座標より、次式

$$Th2' = \{ Xr2 - (Xa1 + Xa2) / 2 \} \tan \theta + Yr1 - Ya2$$
によって垂直統合閾値の変更値 $Th2'$ を計算する。

【0102】実施例18

行統合による文字領域生成までの処理は前記実施例15乃至17と同様であるが、次に文字領域修正部511により統合しきれなかった文字領域の再統合の処理を行なう。この処理の概略フローを図27に示す。

【0103】図27の最初のステップ560は、行統合の段階での処理である。このステップでは、注目行Aに対して一定距離範囲内で下側に水平罫線が存在する場合、生成した文字領域に「下側に水平罫線が存在する」ということを示すマークRを付加する（ステップ562）。例えば図24（a）の文字領域544にマークRが付けられる。

【0104】ステップ562以降が、文字領域再統合の

処理である。まず、新しい文字領域Aを選択する（ステップ562）。この注目文字領域Aとの水平距離及び垂直距離が一定値以下の文字領域をサーチし、見つかったならば、その文字領域がマークR付きの領域であるか調べる（ステップ563）。マークR付き行でなければ、統合を行ない（ステップ564）、次の文字領域のサーチを行なうことになる。

【0105】サーチした文字領域がマークR付き領域である場合、この文字領域の垂直方向の終点座標（Yr2）と文字領域Aの垂直方向の終点座標（Ya2）とが、 $Ya2 > Yr2$ の関係を満たすか判定する（ステップ567）。この関係を満たさないときは統合を行なう（ステップ564）が、関係を満たすときは統合を行わず次の文字領域のサーチに進む。

【0106】ステップ567の条件が満たされるために統合が行なわれない例を図28に示す。文字領域569はマークR付きであるので、注目文字領域A568とは統合されない。このような文字領域は、水平罫線によって区切られた別々の領域と看做することができる。

【0107】図32は後記実施例19の説明のための図である。図32の（a）は入力される文書の例を示す。この文書において、660は縦横罫線で囲まれた表の領域、661はグラフの領域、網掛けされた部分は文字領域（一つ以上の文字あるいは行からなる領域）である。

【0108】従来、この文書の文字領域抽出を行なった場合、図32（b）に示すように、グラフ領域661や表領域660と、それに近接した文字領域とが誤って統合されてしまいやすい。

【0109】後記実施例19によれば、このような文字領域の誤統合を防止することによって、図32（a）の文書に対して図32（c）に示すような領域分割が可能となる。すなわち、後記実施例19によれば、図32（c）に見られるように、グラフ領域661の最下部の水平罫線663と表領域660の最上部の水平罫線664を、文字領域（コラム）の区切りのための架空の水平セパレータとして認識することによって、それを境に上下の文字領域を分離する。

【0110】また、後記実施例20によれば、垂直方向の架空セパレータを認識することによって、垂直方向に近接した図・表領域と文字領域との誤統合を防止する。

【0111】図29は、後記実施例19及び20を説明するためのブロック図である。図29において、601は入力された文書画像データを記憶するための文書画像メモリ、606は入力文書画像の黒連結成分の外接矩形を抽出する矩形抽出部、602は抽出された矩形の情報を記憶する矩形メモリである。607は抽出した矩形を文字矩形と、それ以外のグラフや表などの大きな矩形に分類する矩形分類部である。この分類結果の情報も矩形メモリ602に記憶される。

【0112】608は、図・表矩形より水平罫線または

垂直罫線を抽出する罫線抽出部である。603は抽出された罫線の情報を記憶する罫線メモリである。609は抽出された水平または垂直罫線が水平または垂直の架空セパレータとして妥当であるか判定する架空セパレータ検定部であり、604は架空セパレータの情報を記憶する架空セパレータメモリである。610は文字領域の生成を行なう文字領域生成部であり、文字領域生成のための領域統合の際に架空セパレータを実際のセパレータとして利用する。605は生成された文字領域の情報を記憶する領域メモリである。611は以上の各部を制御する制御部である。

【0113】以下、実施例19及び20を個別に説明する。

【0114】実施例19

図30は架空セパレータを抽出するための処理の概略フローを示す。矩形抽出部606によって抽出された矩形の中から、幅（水平方向の大きさ） W と高さ H が、 $W > LARGEHTH$ かつ $H > LARGEVTH$ の矩形

(1)を選ぶ(ステップ620、621)。これは矩形分類部607の処理である。この矩形(1)は、図32(a)に示した文書の表領域660やグラフ領域661のような文字以外の大きな領域に相当するものである。なお、閾値 $LARGEHTH$ 及び $LARGEVTH$ は、予め与えられた値であるか、あるいは、矩形の高さのヒストグラム等から適応的に決定した値である。

【0115】次に、矩形(1)に対して罫線抽出部608で水平罫線の抽出を行なう(ステップ622)。具体的には、矩形(1)の占める画像の範囲について水平方向にスキャンして黒ランを検出し、閾値 $RUNHTH$ 以上の長さの黒ランを抽出し、この長い黒ランのみの連結成分に外接した矩形(2)を水平罫線として抽出する。この際、矩形(2)の幅 $W1$ と高さ $H1$ も抽出される。抽出された矩形(2)の情報は罫線矩形メモリ603に格納される。前記表領域660のような複数の水平罫線を含む矩形(1)では、複数の矩形(2)が抽出されることになる。閾値 $RUNHTH$ は予め与えた値であるか、あるいは矩形の高さのヒストグラム等から適応的に決定された値である。

【0116】次に、架空セパレータ検定部609において、矩形(1)より抽出された水平罫線矩形(2)について架空水平セパレータとして妥当であるか検定する(ステップ623、624)。次のa~cの条件を全て満足する矩形(2)は架空水平セパレータであると判定され、その情報が架空セパレータメモリ604に格納される。

【0117】a) $W1/W > \text{閾値}$ (例えば0.8)

(ただし、 W は元の矩形(1)の幅)

b) $H1 > \text{閾値} RLHeightTH$

($RLHeightTH$ は、予め与えられた値であるか、あるいは H から自動的に決定される値)

c) 矩形(2)がupperまたはlowerである。

【0118】条件c)について、図31により説明する。図31(a)は矩形(1)としての表領域の一例を示す。(b)はupperとlowerの説明図であり、630~632は水平罫線の矩形(2)を示す。矩形(1)の上辺よりある範囲 $RangeUTH$ 内で最も上にある矩形(2)がupperである。つまり、この例では矩形630がupperである。また、矩形(1)の下辺よりある範囲 $RangeLTH$ 内で最も下にある矩形(2)がlowerである。この例では、矩形632がlowerである。

【0119】文字領域生成部610においては、以上のような処理により抽出された架空水平セパレータを前記実施例15乃至18における水平罫線(水平セパレータ)と同じものとして扱って、前記実施例15乃至18と同様の処理により文字矩形の統合を行なって文字領域を生成し、生成した文字領域の情報を領域メモリ605に格納する。

【0120】実施例20

本実施例では、垂直方向の架空セパレータを抽出する。すなわち、架空セパレータの抽出処理の全体的フローは図23のように示されるが、セパレータの方向の違いに応じてステップ622~625の内容が変更される。

【0121】すなわち、ステップ622において、文字以外の大きな矩形(1)の範囲内の画像を垂直方向にスキャンして黒ランを検出し、閾値 $RUNTH$ 以上の長さの黒ランのみの連結成分に外接する矩形(2)を抽出する。つまり、垂直罫線の矩形を抽出する。

【0122】ステップ623~625で、矩形(1)より抽出された全ての垂直罫線たる矩形(2)に対して、架空垂直セパレータとしての妥当性を検定する。次のd~fの全ての条件を満足する矩形(2)を架空垂直セパレータと判定する。

【0123】d) $H1/H > \text{閾値}$ (例えば0.8)

(ただし、 H は元の矩形(1)の高さ)

e) $W1 > \text{閾値} RLWidthTH$

($RLWidth$ は、予め与えた値であるか、あるいは適応的に決定される値)

f) 矩形(2)がleftまたはrightである。

【0124】条件f)について、図31により説明する。図31(c)はleftとrightの説明図であり、634~636は垂直罫線の矩形(2)を示す。矩形(1)の左辺よりある範囲 $RangeLTH$ 内で最も左にある矩形(2)がleftである。この例では矩形634がleftである。また、矩形(1)の右辺よりある範囲 $RangeRTH$ 内で最も右にある矩形(2)がrightである。この例では、矩形635がrightである。

【0125】文字領域生成部610においては、以上のような処理により抽出された架空垂直セパレータを前記

実施例10乃至14における垂直罫線（垂直セパレータ）と同じものとして扱って、前記実施例10乃至14と同様の処理により文字矩形の統合を行なって文字領域を生成し、生成した文字領域の情報を領域メモリ605に格納する。

【0126】

【発明の効果】請求項1乃至5の発明によれば、様々な文字サイズの文書の画像に対して、文字の矩形と、それ以外の矩形（図表矩形）とを正確に分類できるようになるため、文字矩形の集合たる文字領域の抽出精度が向上する。請求項4または5の発明によれば、書籍や厚手の原稿から読み取られた文書画像に対しても、中折れ部や周辺部に生じる黒領域の影響を排除し、文字矩形、行または文字領域を正確に抽出することができる。

【0127】請求項6乃至10の発明によれば、垂直セパレータまたは水平セパレータで区切られた文字領域（コラム）の誤統合を防止し、それぞれの文字領域を正しく分離できる。また、処理内容も簡便であって処理の高速化が容易であり、さらに処理のために必要なメモリ量も少なくて済む。

【0128】請求項11または12の発明によれば、表やグラフの領域と、それに接近した文字領域との誤統合を防止し、領域分割の精度を大幅に向上できる。

【図面の簡単な説明】

【図1】実施例1乃至実施例6の説明のためのブロック図である。

【図2】（a）乃至（d）処理概要の説明のための図である。

【図3】実施例1の矩形抽出から標準文字サイズ決定までの処理フロー図である。

【図4】（a）及び（b）矩形の高さのヒストグラムの例を示す図である。

【図5】実施例2における文字矩形統合の説明図である。

【図6】文字矩形の統合により抽出された行を示す図である。

【図7】実施例3における文字矩形統合による行抽出の説明図である。

【図8】実施例5における行統合による文字領域抽出の説明図である。

【図9】実施例5における行統合による文字領域抽出の説明図である。

【図10】実施例6における大きな文字領域の統合の説明図である。

【図11】実施例6における小さな文字領域の統合の説明図である。

【図12】実施例7乃至実施例9の説明のためのブロック図である。

【図13】実施例7におけるノイズ矩形除去処理のフロー図である。

【図14】実施例8及び実施例9におけるノイズ矩形除去の説明図である。

【図15】実施例10乃至実施例14の説明のためのブロック図である。

【図16】実施例16における行生成処理のフロー図である。

【図17】（a）及び（b）実施例10における行統合と垂直罫線との関係の説明図である。

【図18】実施例11における垂直罫線までの距離の求め方の説明図である。

【図19】（a）及び（b）実施例12における垂直罫線までの距離の求め方の説明図である。

【図20】実施例13における行統合処理のフロー図である。

【図21】（a）及び（b）実施例13における行統合の説明図である。

【図22】実施例15乃至実施例18の説明のためのブロック図である。

【図23】実施例15における行統合処理のフロー図である。

【図24】（a）及び（b）実施例15における行統合と水平罫線との関係の説明図である。

【図25】実施例16における水平罫線までの距離の求め方の説明図である。

【図26】（a）及び（b）実施例17における水平罫線までの距離の求め方の説明図である。

【図27】実施例18における文字領域再統合処理のフロー図である。

【図28】実施例18における文字領域再統合の説明図である。

【図29】実施例19及び実施例20の説明のためのブロック図である。

【図30】実施例19における架空水平セパレータの抽出処理のフロー図である。

【図31】（a）乃至（c）実施例19及び実施例20における架空セパレータ抽出の説明図である。

【図32】（a）乃至（c）実施例19による文字領域の誤統合の防止の説明図である。

【符号の説明】

1, 301, 501, 601 文書画像メモリ

2, 302, 502, 602 矩形メモリ

3 ヒストグラムメモリ

4 標準文字サイズ保持メモリ

5, 306, 506, 606 矩形抽出部

6 ヒストグラム生成部

7 標準文字サイズ決定部

8, 307, 507, 607 矩形分類部

9, 309, 509 行抽出部

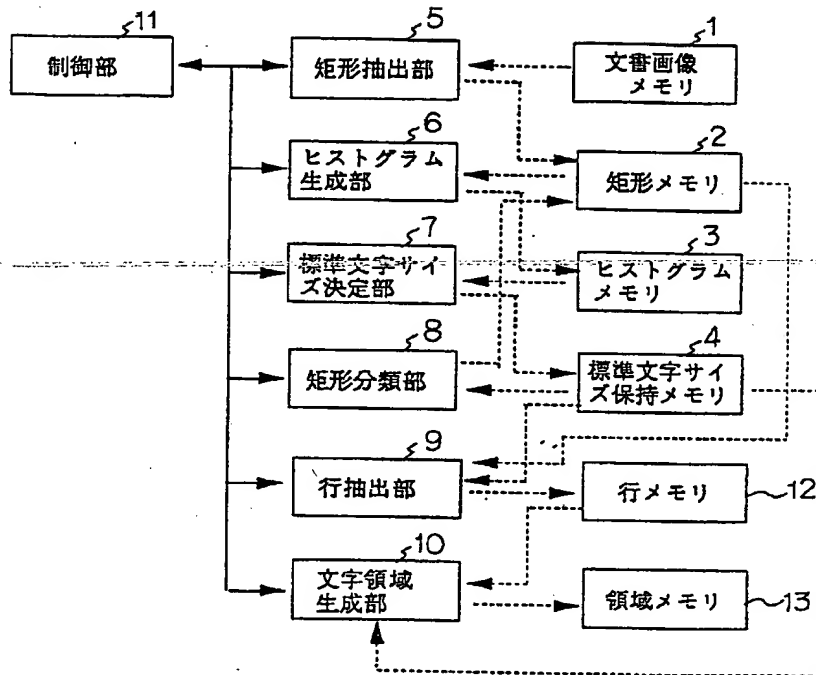
10, 310, 510 文字領域生成部

11, 312, 512, 611 制御部

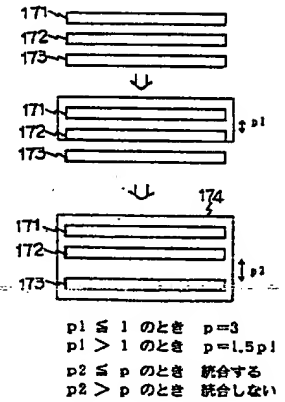
12, 304, 503 行メモリ
 13, 305, 505, 605 領域メモリ
 201 ノイズ矩形除去部
 303, 504 統合閾値メモリ
 308 垂直罫線識別部
 311, 211 文字領域修正部

508 水平罫線識別部
 603 罫線メモリ
 604 架空セパレータメモリ
 608 罫線抽出部
 609 架空セパレータ検定部
 610 文字領域生成部

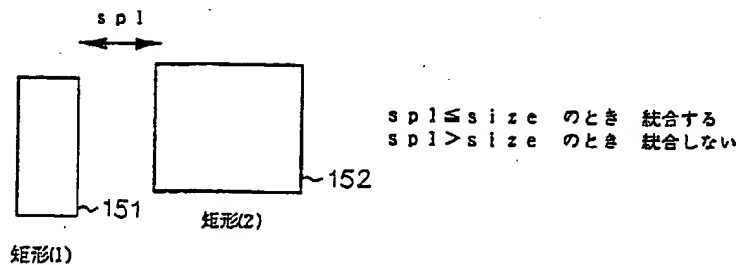
【図1】



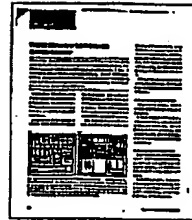
【図8】



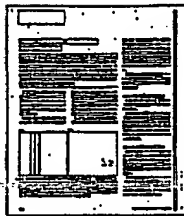
【図5】



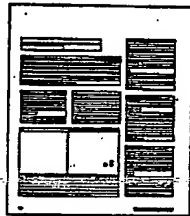
【図2】



(a) 入力原稿



(b) 矩形データ

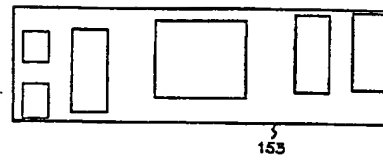


(c) 行/領域データ

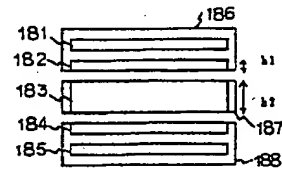
2個から成る

(d) 矩形データ (拡大)

【図6】

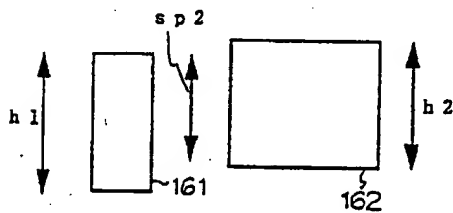


【図9】



$h2/h1 \leq 4/5$ のとき 統合しない
 $h2/h1 < 4/5$ のとき 統合する

【図7】

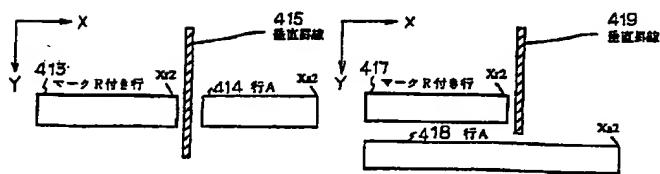


$h1$: 矩形1の高さ
 $h2$: 矩形2の高さ

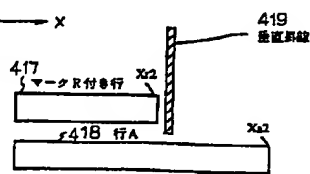
$h1 \leq h2$ のとき $h = h1$
 $h1 > h2$ のとき $h = h2$

$sp2 \geq h/2$ のとき 統合する
 $sp2 < h/2$ のとき 統合しない

【図21】

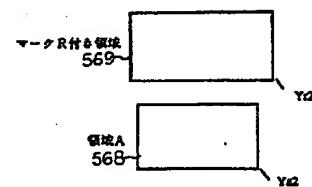


(a) 行の統合を行わない例

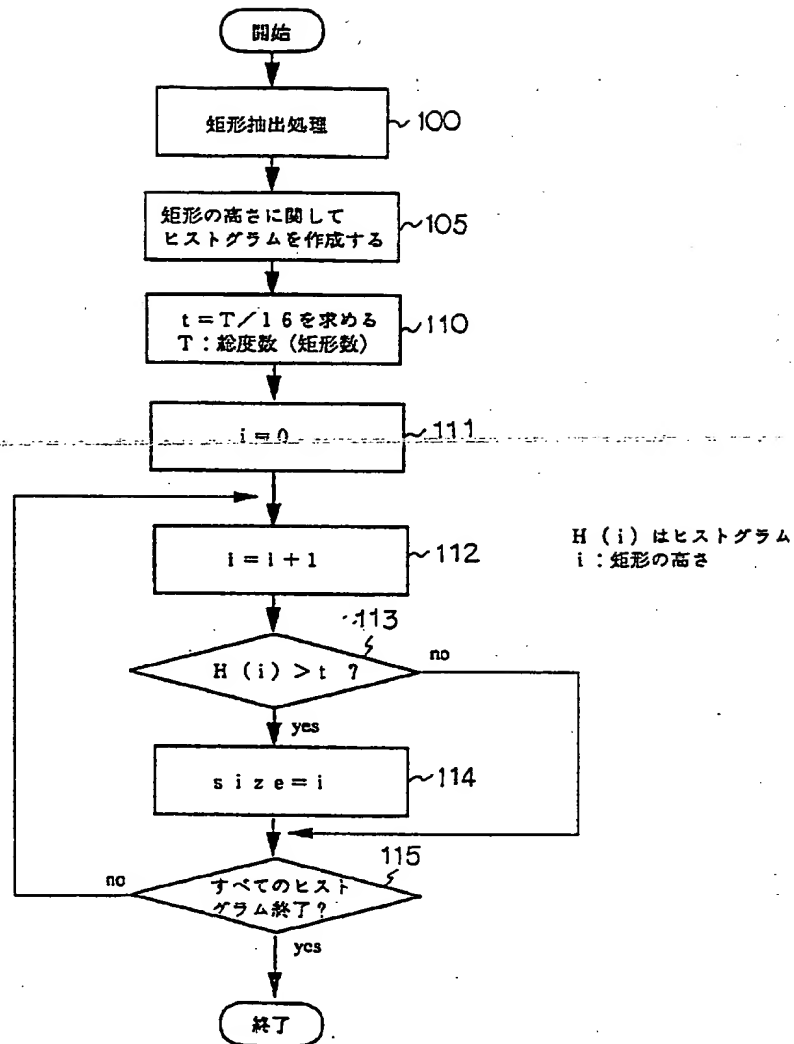


(b) 行の統合を行わない例

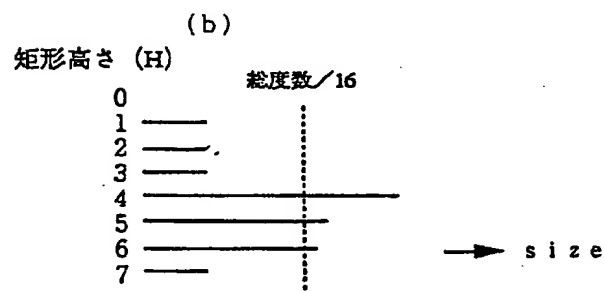
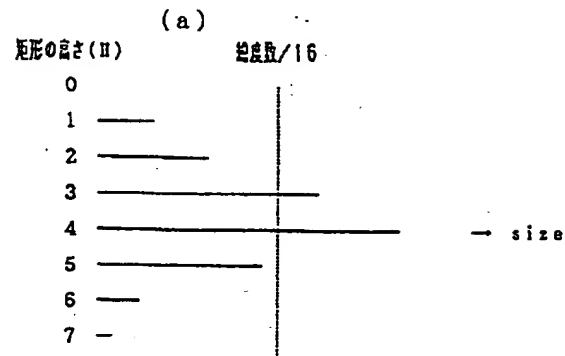
【図28】



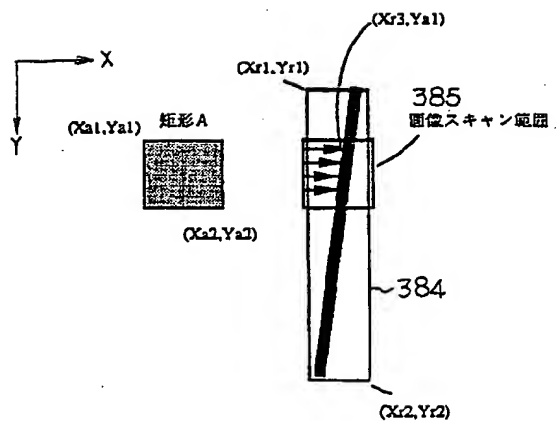
【図3】



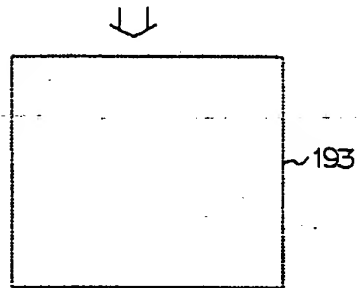
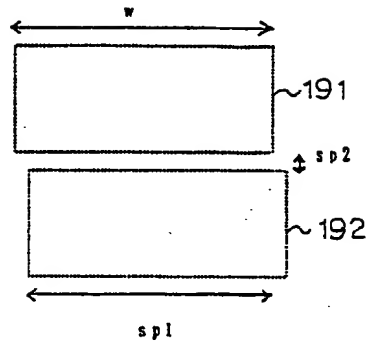
【図4】



【図18】



【図10】

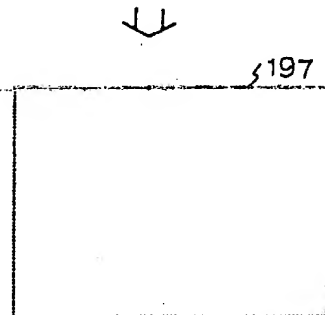
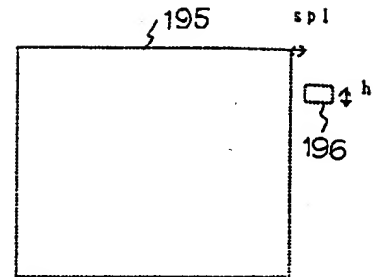


$$sp1 \geq 4/5 w$$

$$sp2 < size$$

(大きい領域同士のと看)

【図11】

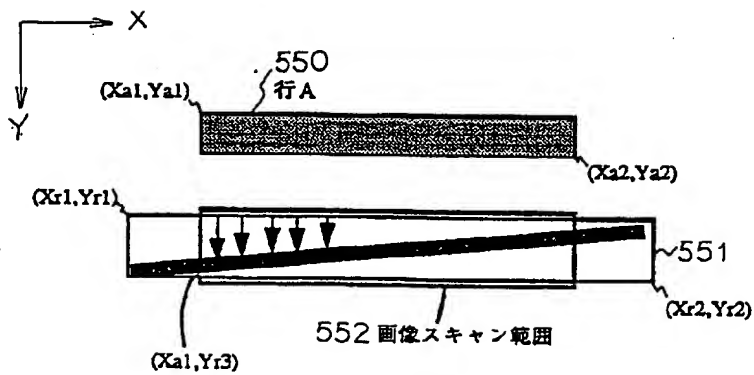


$$sp1 \leq 2size$$

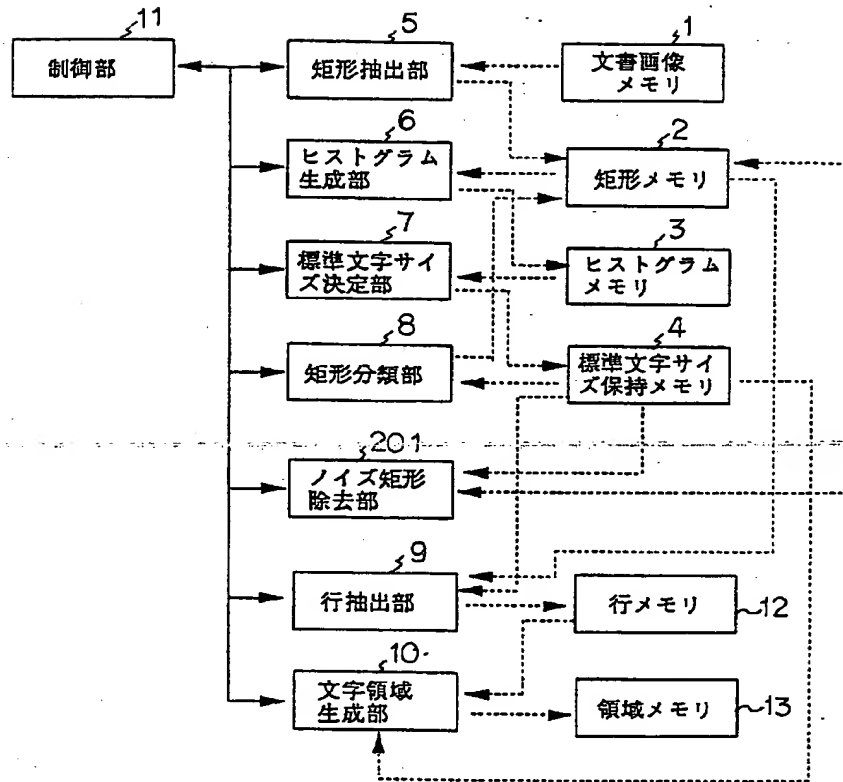
重なりが $h/2$ 以上

(小さい領域のと看)

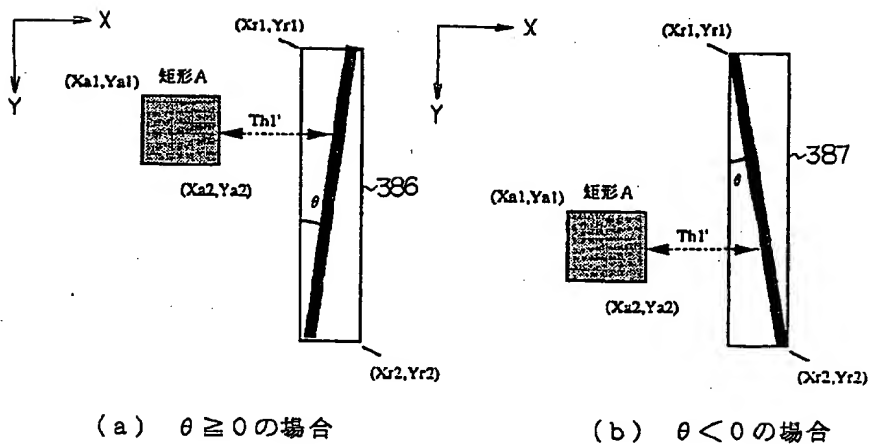
【図25】



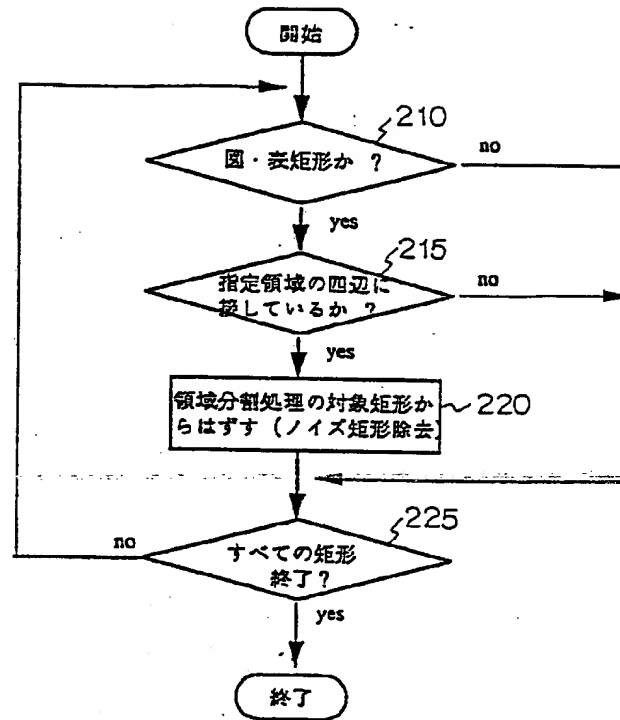
【図12】



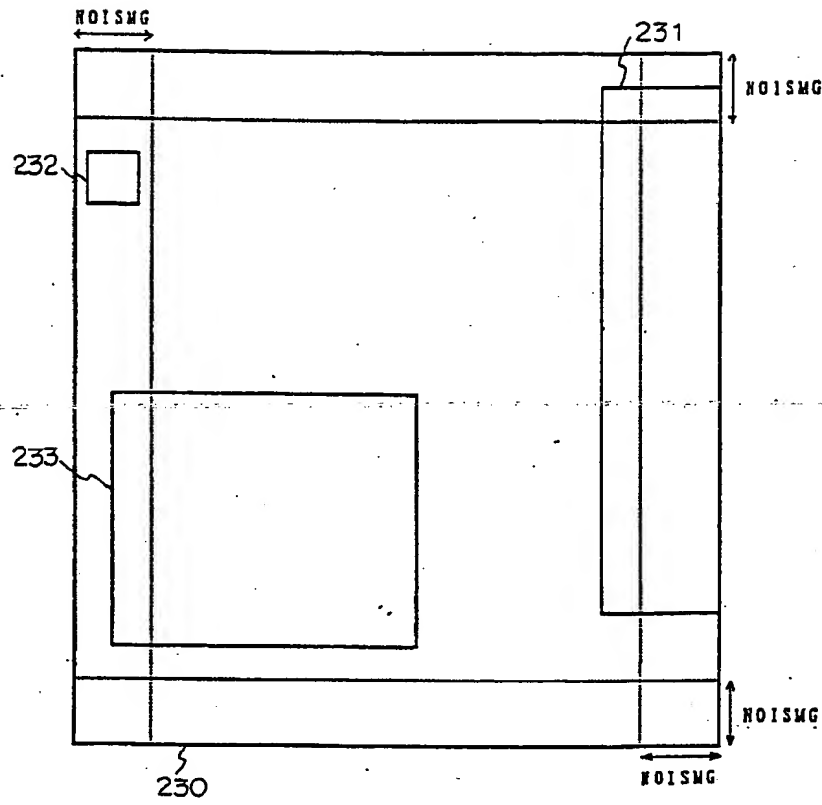
【図19】



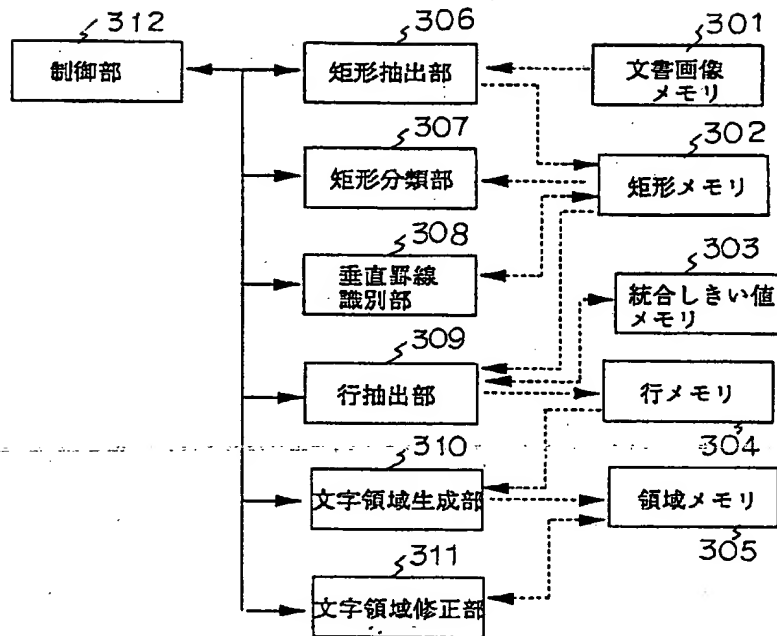
【図13】



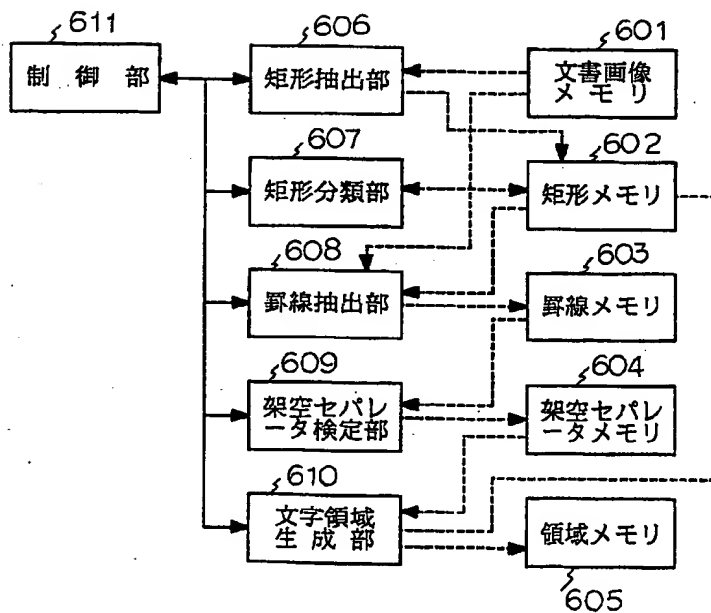
【図14】



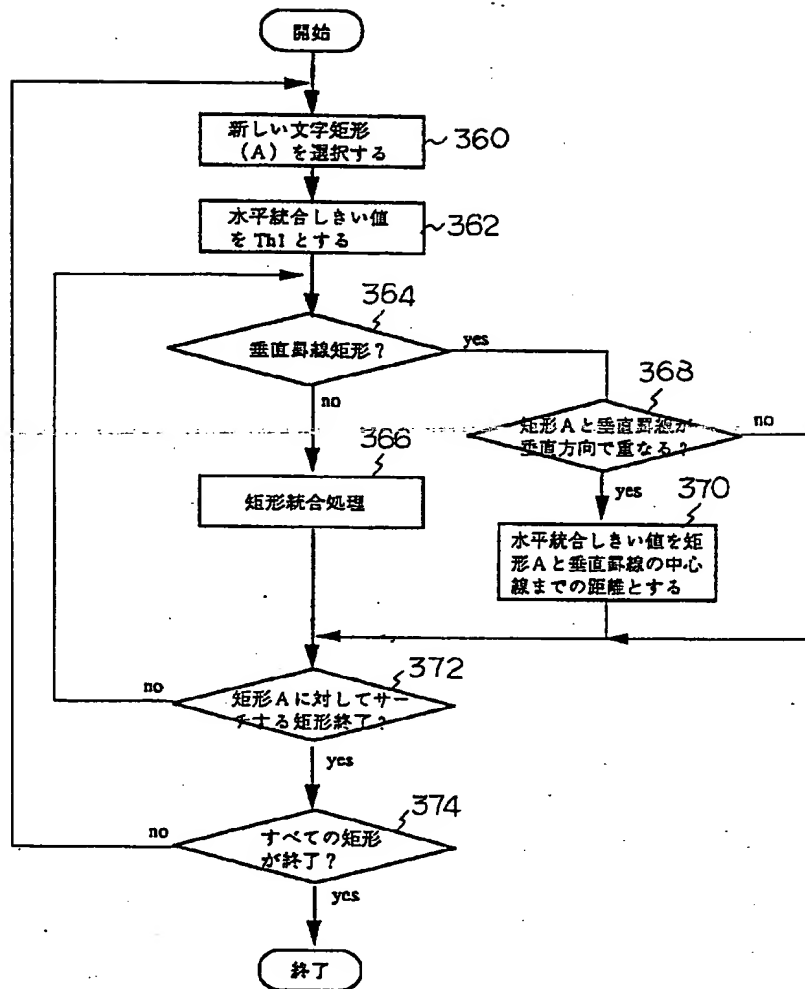
【図15】



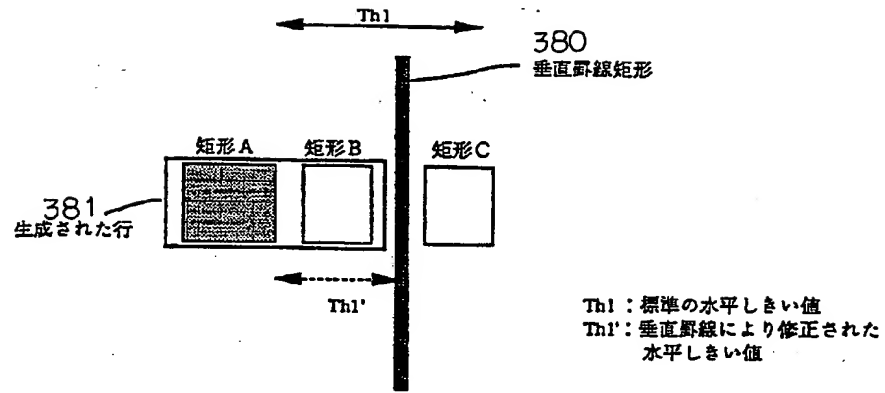
【図29】



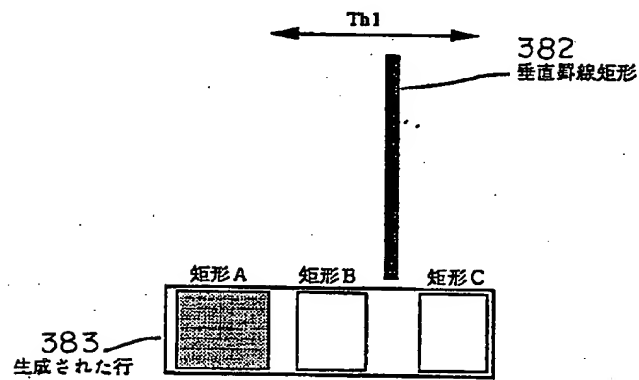
【図16】



【図17】

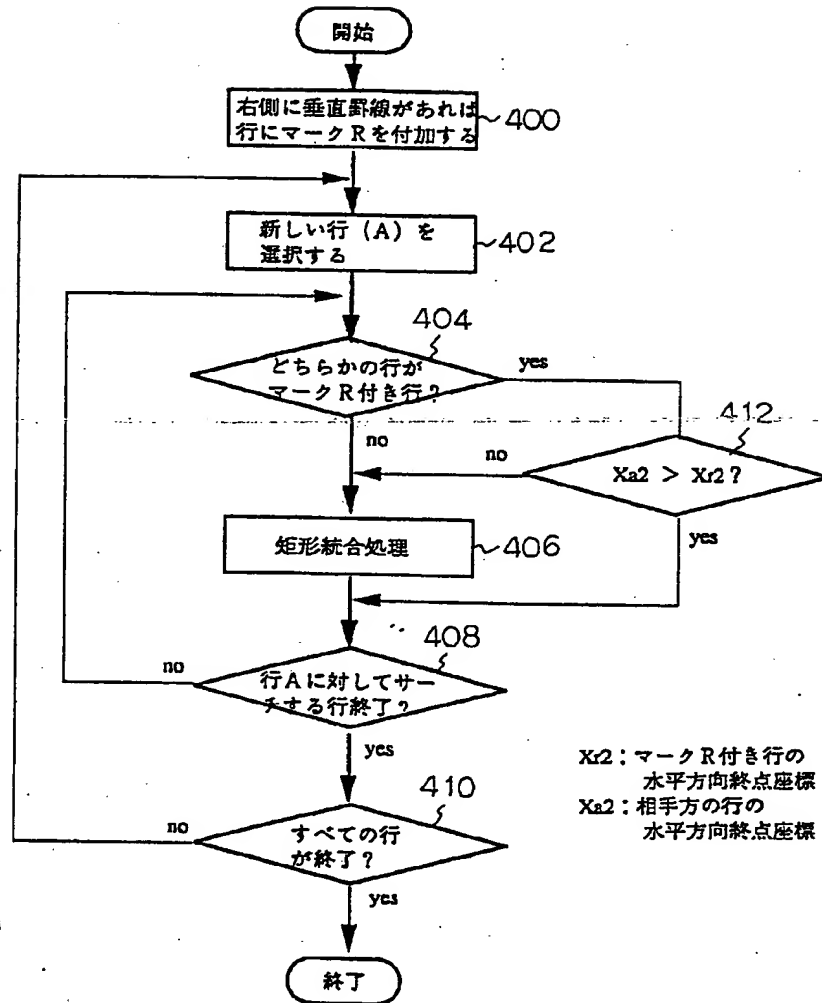


(a) 垂直罫線と水平統合しきい値

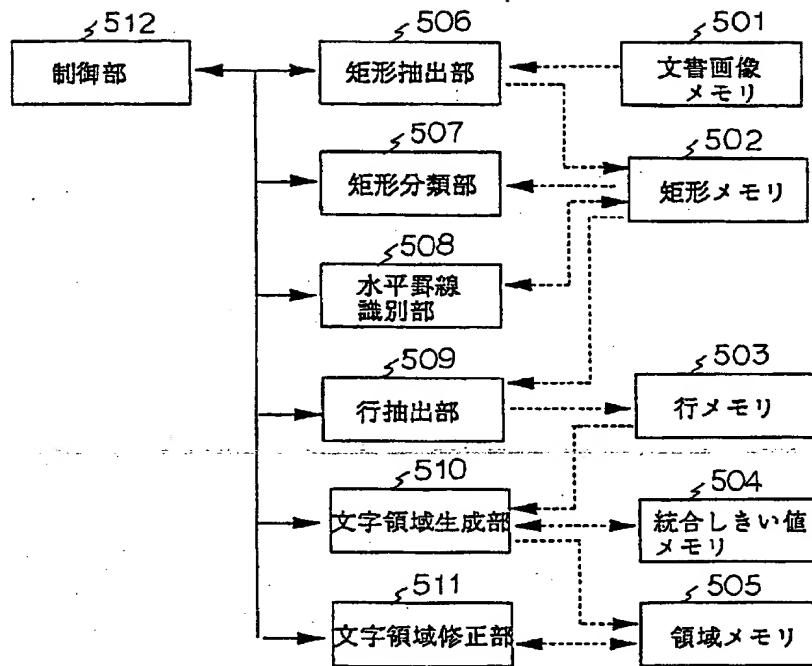


(b) 垂直罫線と矩形Aが垂直方向に重ならない場合

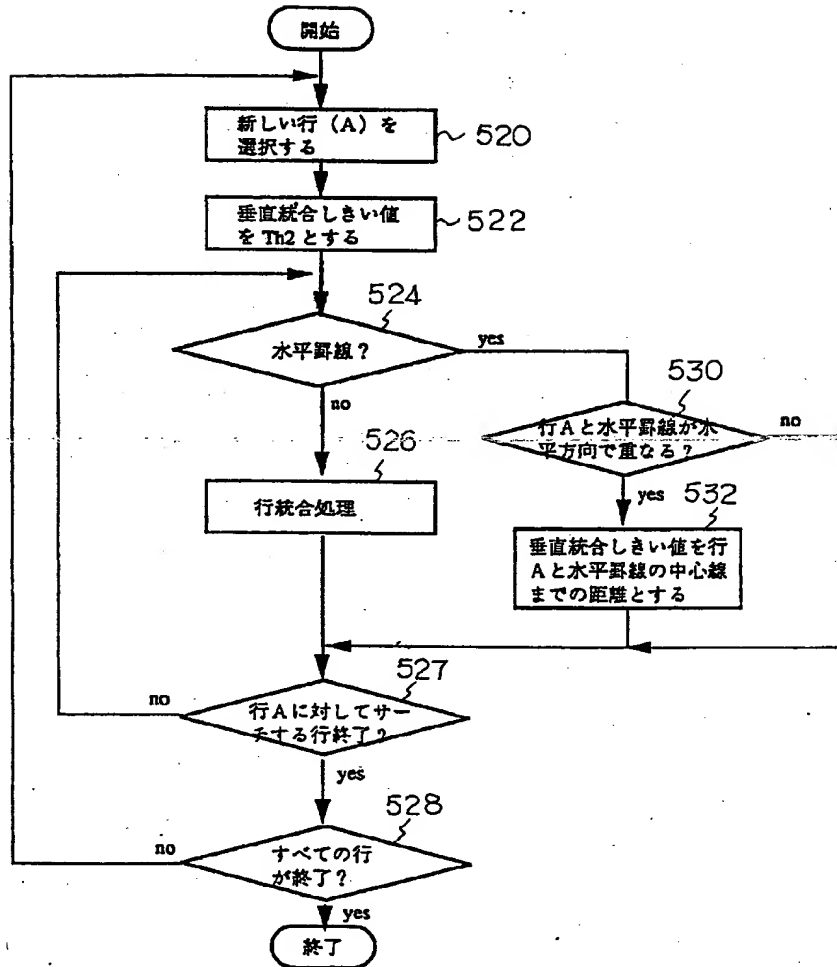
【図20】



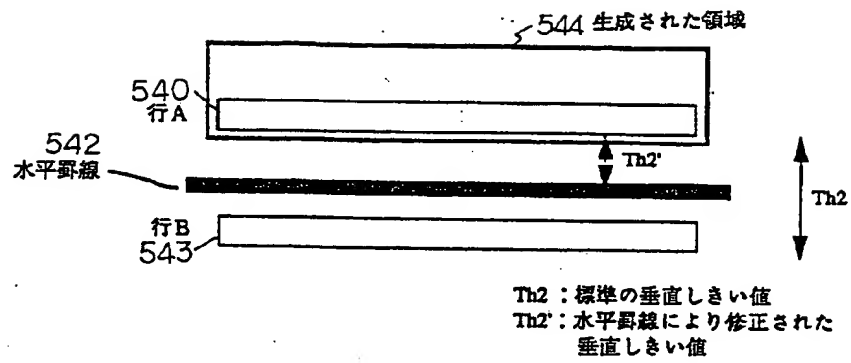
【図22】



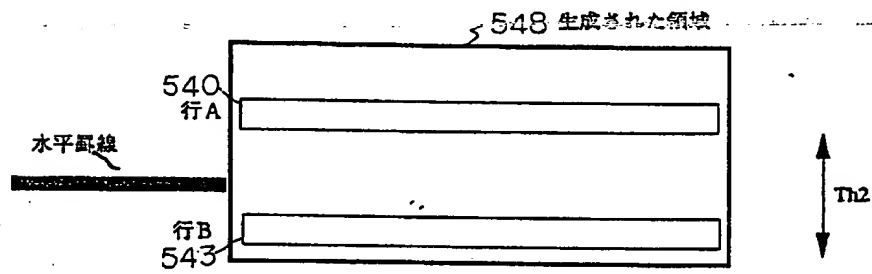
【図23】



【図24】

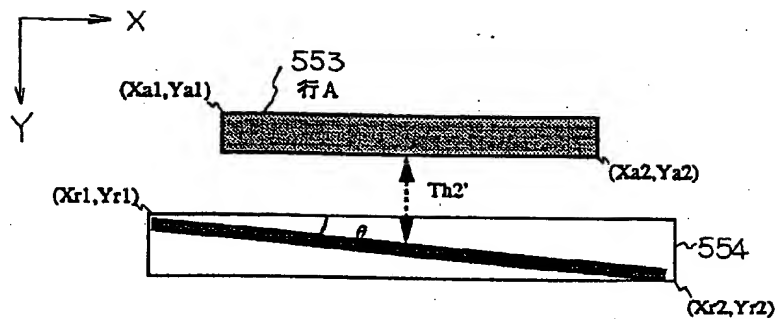
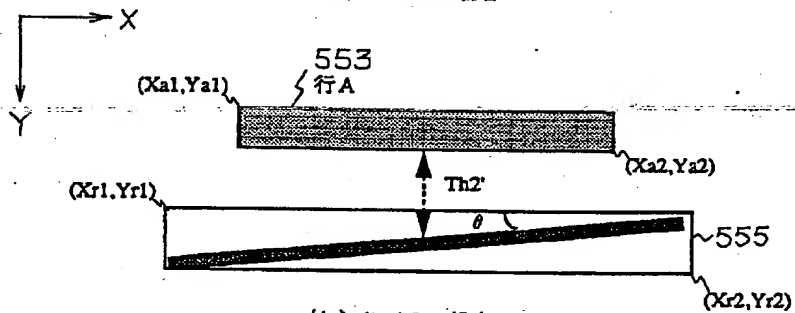


(a) 水平罫線によりコラムが分割される場合

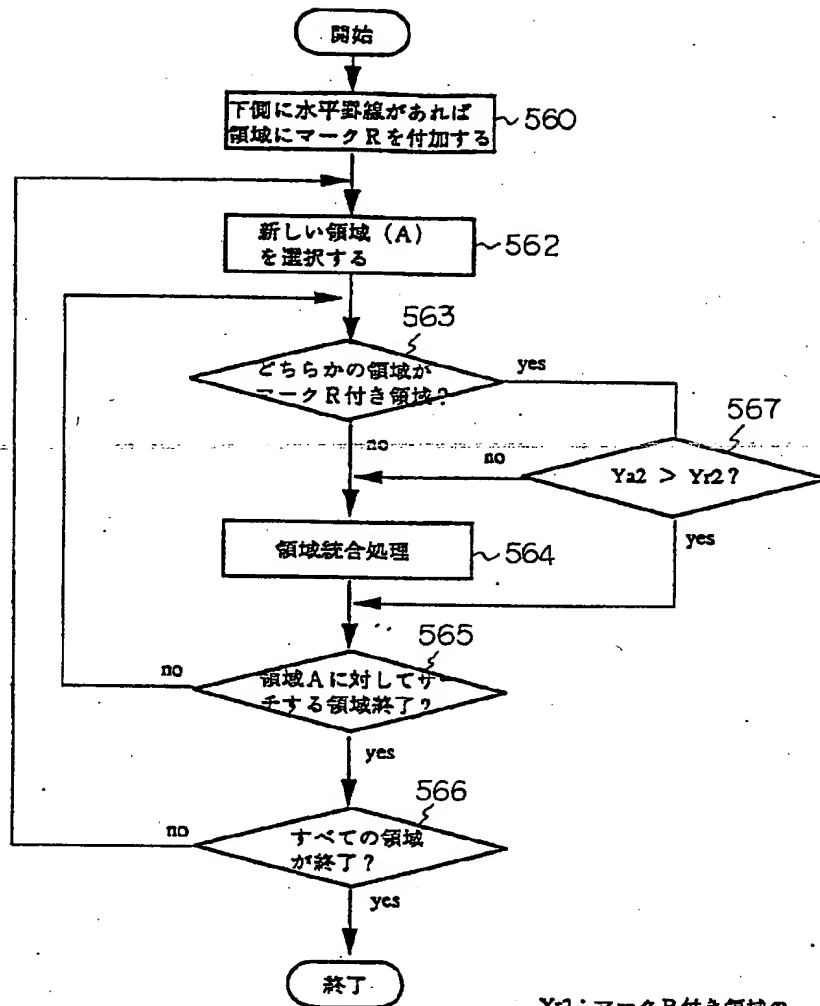


(b) 水平罫線と行Aが水平方向に重ならない場合

【図26】

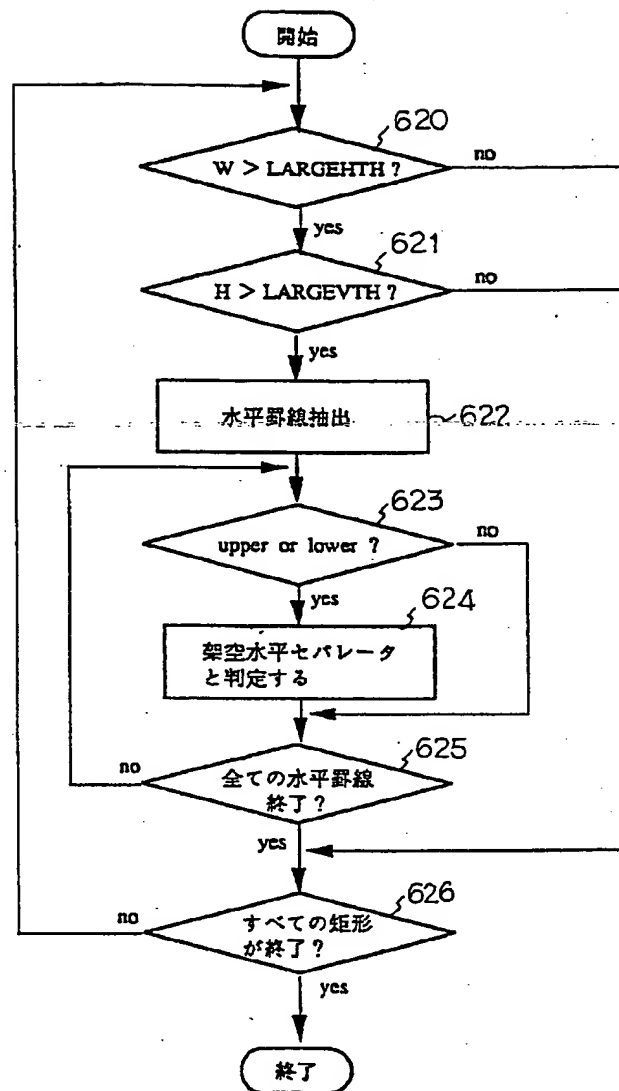
(a) $\theta \geq 0$ の場合(b) $\theta < 0$ の場合

【図27】

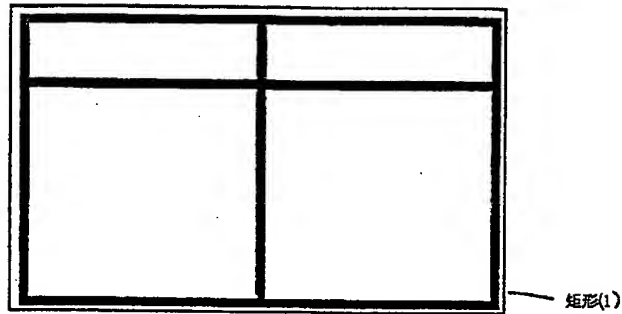


Yr2: マークR付き領域の
垂直方向終点座標
Ya2: 相手方の領域の
垂直方向終点座標

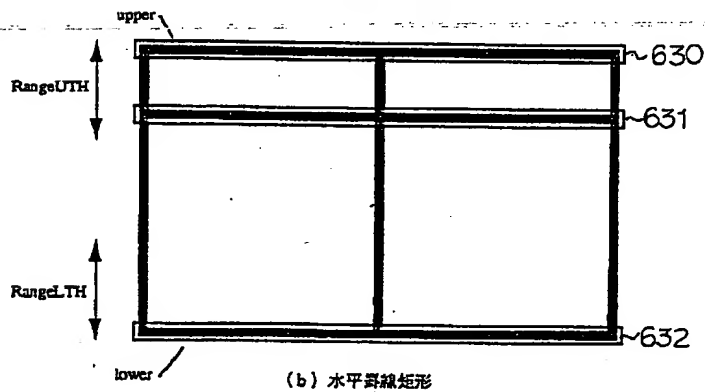
【図30】



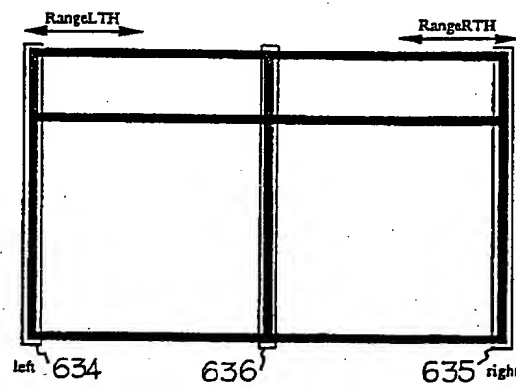
【図31】



(a) その他矩形候補 (矩形1)

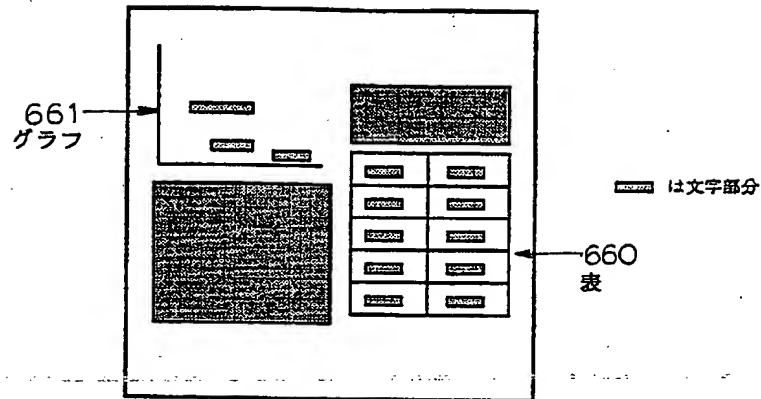


(b) 水平割線矩形

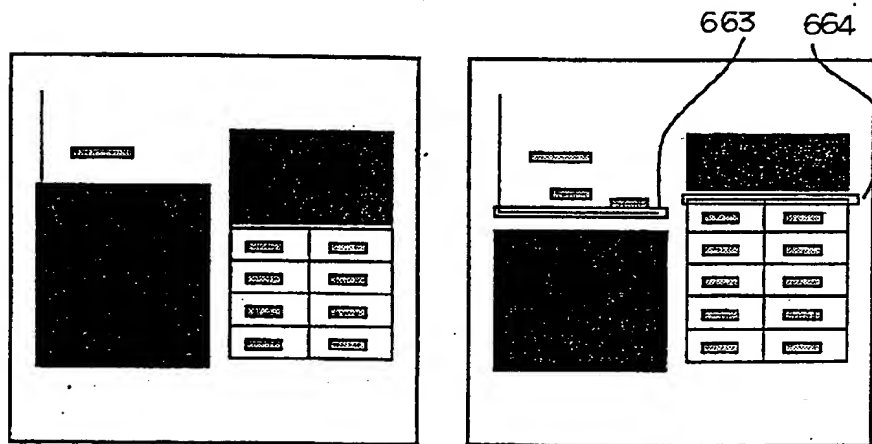


(c) 垂直割線矩形

【図32】



(a) 入力された文書



663 は文字部分として抽出された領域

(b) 文字部の統合が失敗した例

(c) 本発明による領域分割結果